PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : H04L 12/56, 12/64, H04Q 11/04 — A1

(11) International Publication Number: WO 00/38375

(43) International Publication Date: 29 June 2000 (29.06.00)

(21) International Application Number: PCT/GB99/03748

(22) International Filing Date: 10 November 1999 (10.11.99)

(30) Priority Data:
9828144.7 22 December 1998 (22.12.98) GB

(71) Applicant *(for all designated States except US)*: POWER X LIMITED [GB/GB]; Stafford Court, 145 Washway Road, Sale, Cheshire M33 7PE (GB).

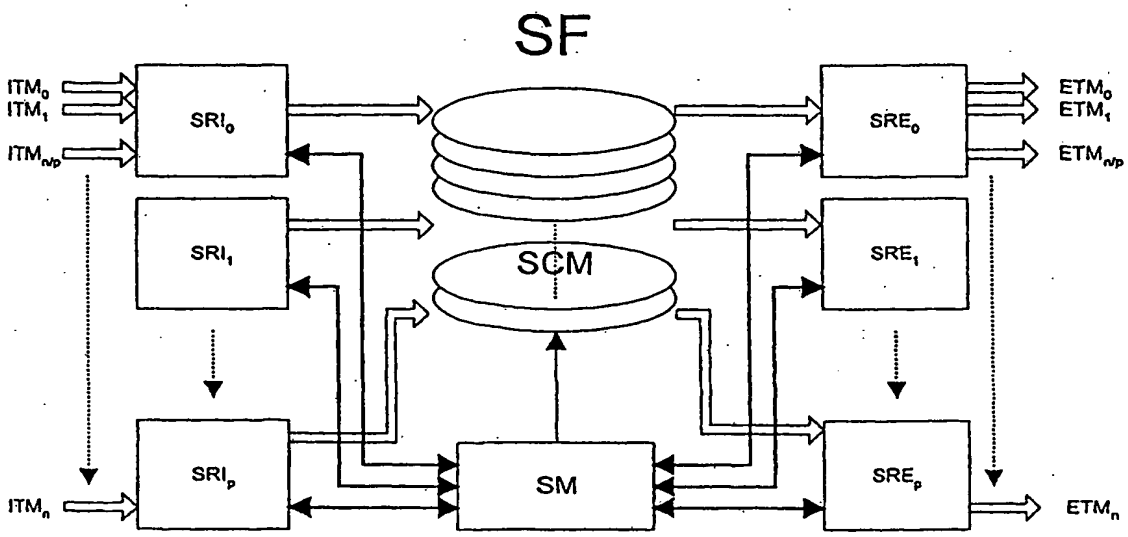
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: JOHNSON, Ian, David [GB/GB]; 11 Seel Street, Moseley, Manchester OL5 0EW (GB). COLLINS, Michael, Patrick, Robert [GB/GB]; 53 Brompton Road, Rusholme, Manchester M14 7QA (GB). HOWARTH, Paul [GB/GB]; 14 Badby Close, Ancoats, Manchester M4 7EY (GB).

(74) Agents: MCNEIGHT, David, Leslie et al.; McNeight & Lawrence, Regent House, Heaton Lane, Stockport, Cheshire SK4 1BS (GB).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: DATA SWITCHING METHOD AND APPARATUS



(57) Abstract

A data switch for handling packets of information switch comprises input traffic managers, ingress routers, a memoryless cyclic switch fabric, egress routers and output traffic managers all acting under the control of a switch controller. Each ingress router includes a set of virtual output buffers one for each output traffic manager and each message priority. Each data packet or cell as it arrives is examined to identify the output traffic manager address and its message priority. The switch controller uses a first arbitration and selection process to schedule the passage of the next cell across the switch fabric which the ingress router uses a second arbitration and selection process to select the appropriate virtual output queue for use in the switch fabric transfer.

DATA SWITCHING METHOD AND APPARATUS

This invention relates to a method of data switching which takes application data from numerous input sources and routes it to numerous destination outputs and to apparatus for performing such switching.

In a generalisation of such a concept, data arriving on input ports is routed via a non-blocking cross bar switch to output ports. For an input N to transfer data to an output M the switch establishes a 'connection' between N and M. The connection generally remains for the duration of the data transfer at which point it may be broken and the output allowed to be connected to another input. Data is typically transferred in 'cells'.

Because there are numerous inputs competing for numerous output ports the possibility of contention occurs. The output port can be considered to be a resource that must be shared amongst multiple inputs. This means that a particular input may not be able to connect to a particular output because that output is already in use i.e. is already connected to another port. It is also possible that more than one input may be requesting a connection to the same output. In either case the result is the need for the cells or data products to queued (buffered) until the relevant resource becomes available.

Cells can be stored in several areas in the switch; the input, the output and centrally. Most switches use a combination of all three. It is generally considered that output buffering provides the most efficient way for handling traffic shaping i.e. the profile of the release of cells from the switch. However, output buffering places severe requirements on the actual storage device used to create the buffer. This is because the buffer is shared amongst multiple inputs

the data switch to the appropriate output traffic controller in accordance with a second arbitration process.

According to a second aspect of the invention there is provided a data switch for handling packets of information comprising input traffic controllers, ingress routers, a memoryless cyclic switch fabric, egress routers and output traffic controllers all under the control of a switch controller and interconnected such that each input line connected to the data switch is terminated on a traffic controller arranged to convert the input line protocol information packets into fixed length cells having a header defining the data switch destination router and output traffic controller together with message priority information arranged such that each ingress router serves a group of traffic controllers characterised in that the ingress router includes a set of input buffers one for each input line and a set of virtual output queue buffers, one for each output traffic controller connected to the data switch, and in which on the arrival of a cell from a traffic controller the ingress router examines the cell header and places it in the appropriate virtual output queue and generates a request for transfer message consisting of the destination traffic controller address and a message priority code which is passed to the data switch controller, the switch controller schedules the passage of the cells across the switch fabric by interconnecting a specific ingress router to a specific egress router for each switch fabric cycle in accordance with a first arbitration process and the ingress router selects from the appropriate virtual output queue the cell at the head of the queue for passage across the data switch to the appropriate output traffic controller in accordance with a second arbitration process.

The invention together with its various features will be more readily understood from the following description which should be read in conjunction with the accompanying drawings, in which:-

reasons. Input buffering requires smaller buffers, which can have relatively low performance and therefore be cheaper.

When cells are queued at the input there is the possibility of contention arising through the phenomena of Head Of Line (HOL) blocking. This generally occurs when First In First Out (FIFO) queue mechanisms are used. The FIFO queues the cell at the head of the queue and this is the only one that can be chosen for delivery through the switch. Now, consider the case where an input port has three cells cl, c2, c3 stored such that c1 is at the head of the queue with c2 stored next and c3 last with cell c1 destined for port N and cell c2 destined for port N+1. Now port N is already connected to port N-1 therefore c1 cannot be switched, however port N+1 is unconnected and therefore c2 could actually be delivered. However, c2 cannot get out of the FIFO because it is blocked by the HOL i.e. c1. An intelligent approach to the solution of HOL blocking is the concept of Virtual Output Queues (VOQ). Using VOQs the cells are separated out at the input into queues which map directly to their required output destination. They can therefore be effectively described at being output queues, which are held at the input i.e. virtual Output Queues. Since the cells are now separated out in terms of their output destination they can no longer be blocked by the HOL phenomena.

There is also the question of Quality Of Service (QoS) to address. Different input sources have different requirements in terms of how their data should be delivered. For example voice data must be guaranteed to a very tightly controlled delivery service whereas the handling of computer data can be more relaxed. To accommodate these requirements the concept of priority can be used. Data is given a level of priority, which changes the way the switch deals with it. For example consider two cells in different VOQs cl and c2 which are both

Referring now to Figure 2, the main feature is the data switch SW. Inputs are provided to the switch from ingress traffic manager units ITM_0 to ITM_n. Each ingress traffic manager may have one or more input line end devices (ILE) connected to it. Outputs from the switch SW are connected by way of egress traffic manager units ETM_0 to ETM_n to egress line end devices (ELE).

The traffic manager units (ITM and ETM) provide the protocol-specific processing in the switch, such as congestion buffering, ingress traffic policing, address translation (ingress and egress) and routing (ingress), traffic shaping (ingress or egress), collection of traffic statistics and line level diagnostics. There may also be some segmentation and re-assembly functionality within a traffic manager unit. The line end devices (ILE and ELE) are full-duplex devices and provide the switch port physical interfaces. Typically, line end devices will be operated in synchronous transfer mode, ranging from OC-3 to OC-48 rates or 10/100 and Gigabit Ethernet.

The switch SW provides the application independent, loss-less transport of data between the traffic managers based on routing information provided by the traffic managers and the connection allocation policies determined by the switch control SC. This controls the global functions of the switch such as connection management, switch level diagnostics, statistics collection and redundancy management.

The switching system just described is based on an input-queued non-blocking crossbar architecture. A combination of adequate buffering, hierarchic flow control, and distributed scheduling and arbitration processes ensure loss-less, efficient, and high performance switching capabilities. It should be noted that the ingress and egress functions are shown separately on either side of the drawing.

switching matrix SCM. The controller SM selects an optimal combination of connections to establish in the matrix SCM once per switching cycle. The selection can be postponed by one (or more) backpressure broadcast requests that are satisfied in a round-robin fashion before allowing normal operation to resume. The arbiter also uses a probabilistic work-conserving algorithm to allocate bandwidth in the switching matrix to each priority according to information defined by the external system controller.

The switching matrix SCM itself consist of a number of memory-less, non-blocking matrix planes SCM1 - N and a number of embedded serial transceivers to interface to the routers. The number of matrix planes in a particular switch depends on the core throughput required across the matrix. The core throughput will be greater than the aggregate of the external interfaces to allow for inter-router communication, core header overheads and maximal connections during the arbitration cycles. The device is packaged with two planes of sixteen ports, which can be configured to provide an alternative number of planes/ports. The multiple serial links that comprise the data path between the router and switching matrix are switched simultaneously and therefore act as a single full duplex fat pipe of 8Gbps. The switching matrix has the novel feature that it can be configured as a 'NxN' port crossbar device where N can be 4, 8 or 16. This feature can increase the number of planes per package and therefore allows a wide range of systems to be realised cost effectively. For example using the first generation chip set systems of less than 20Gbs up to 80Gbs can be easily configured.

Underlying the management of the system is the fabric management interface FMI, which provides an external orthogonal interface into all of the system devices. This level of management provides read/write access to a chosen subset of important registers and RAMs while the device is functioning normally,

Error checking routines are automatically performed during system initialisation. The FMI protocol includes parity in all of its messages.

b) Correction. If an error is detected in a tensor, because either the data is faulty or the tensor has been misrouted, the system will not correct the error. The tensor is discarded and it is left to a higher level of protocol to carry out any necessary corrective actions. Where errors are detected on certain control interfaces, retries are attempted without any external intervention in order to distinguish between a transient and permanent failure. The fault is reported via the FMI in either case.

c) Containment. The principle of containment is to limit the effect of an error and, as far as possible, continue normal operation. For example if a fault is detected in a particular tensor, that tensor is discarded but the system carries on operating normally. Similarly, if a permanent fault is detected that affects one traffic manager unit or router, that part of the system is disabled whilst the rest of the system continues without a break in service. This may require system management assistance. If redundancy were employed in the system, then at this point the standby device(s) would become operational.

d) Reporting. All faults which allow the reporting infrastructure to continue functioning are logged and reported to the diagnostic system. The device primary status register has a mechanism for reporting different classes of fault separately, so that any necessary action can be quickly determined.

e) Monitoring. In addition to error monitoring, the system contains logs to collect performance monitoring and statistics information. These can be dynamically accessed.

The cell is transferred through the memory-less switching matrix SCM and into a buffer in the egress router SRE.

As shown at step 6, there is one egress buffer per egress traffic manager ETM and arriving cells are examined and placed in the appropriate traffic manager queue in the egress router SRE.

Finally, at step 7, the cell is transferred to the egress traffic manager EME over the standard interface CSIX and, where necessary, re-assembled into a packet before onward transmission.

The transfer of data through the system is packaged in cells termed tensors. An arbitration cycle transfers one tensor per router through the switching matric SCM. Each tensor consists of 6 or 8 vectors. A vector consists of one byte per plane of the switching matrix and is transferred through it in one system clock cycle. The sizes of the vector and tensor for a particular application are determined by the bandwidth required in the fabric and the most appropriate cell size. The following sections show the typical packaging of the data as it flows through the system for ATM and Ethernet.

As shown in Figure 5a, illustrating the ASTM application, payload cells P containing fifty three bytes of data arriving from an ingress traffic manager ITM across the interface CSIX are re-packaged into 60-byte tensors (6 vectors of 10 bytes). The ingress router analyses the CSIX header UH and wraps the CSIX packet with the core header CH to create a 60-byte tensor UCT in an ingress queue. When the controller SM grants the required connection the tensor passes through the switching matrix SM in one switch cycle to the egress router which writes the unicast tensor UT into the egress queue indicated in the core header. When the tensor reaches the head of the egress queue, the core header is stripped off and the remaining CSIX packet is sent to the egress traffic manager.

to the routers SR and the controller SCM into the switching matrix SC port. The diagram also shows how information on channel, link bandwidth allocation and switch efficiency, queue status, backpressure and traffic congestion management is handled by the referenced arrows.

The controller SM provides the overall control function of the system. When the routers request connections from the controller, they identify their requested switching matrix connection by switch port and priority. The controller then selects combinations of connections in the switching matrix to make best use of the matrix connectivity and to provide fair service to the routers. This is accomplished by using an arbitration mechanism. The controller SM can also enforce pseudo-static bandwidth allocation across the priorities and ingress/egress switch port combinations. For example, an external system controller can guarantee a proportion of the available bandwidth to each of the priorities and to specific connections. Unused allocations will be fairly shared between other priorities and connections.

The controller SM also has a 'best effort' mechanism to dynamically bias the arbitration in favour of long queues for applications that do not require strict bandwidth enforcement.

The routers provide an aggregation function for multiple traffic managers into a single switch port. When the controller SM grants a connection to a particular egress switch port through a particular priority, the appropriate router must choose one of up to eight unicast and one multicast traffic manager queues to service. This is accomplished through a weighted round-robin mechanism, which can select a queue based on a combination of ingress queue length. These may allow for favouring of long queues over shorter ones, or allows traffic

requesting egress router .The egress router then sends one vector's worth (10 bytes) of egress buffer status through the matrix to the ingress routers. The controller then continues the interrupted cycles. In the event of several egress routers simultaneously requesting a backpressure broadcast, the controller will satisfy all the requests in a simple round-robin manner before resuming normal service. The latency introduced in the backpressure mechanism due to this contention does not affect the egress buffering since during this period a router will only be receiving backpressure data from other routers, which does not need to be queued.

An egress router will aggregate the threshold transitions from all its egress queues, which have occurred during a switch cycle into one backpressure broadcast so that the maximum number of backpressure broadcasts between two tensors, is limited to the number of routers. When an ingress router ingress receives a backpressure broadcast vector of the form shown in Figure 8, it uses it to update the ingress queue weightings as appropriate.

Two modes of backpressure signalling between egress and ingress routers are supported, namely start/stop and multi-state signalling. Multi-state signalling allows the egress router to signal the multi-bit state of all its queues (1 byte per queue). This multi-state backpressure signalling coupled with weighted-round-robin scheduling in the ingress routers minimises the probability of egress queues being full, which is significant when attempting to forward multicast or broadcast traffic in a heavily utilised switch.

The ingress router signals stop/start backpressure to the ingress traffic managers via the CSIX interface. This provides a 16-bit backpressure signal to allow the ingress router to identify the ingress queue to which the signal relates.

multicast performance. These are: 1. multi-state backpressure from the egress router, which reduces the probability of egress queues being full, and 2. increasing the weighting of the multicast ingress queues in the weighted-round-robin scheduler when they have been blocked to increase their chances of being scheduled when the block clears. To avoid multicast (and broadcast) being blocked by off-line egress ports, the backpressure signals can be individually masked out by an external system controller via the Fabric Management Interface (FMI).

The requirement for wire-speed broadcast (benchmarking) is met by having a single on-chip broadcast queue in each egress router. When the controller schedules a broadcast connection, the tensor will be routed in the switching matrix to all routers in parallel, thus avoiding any ingress congestion (no tensor replication at ingress). Broadcast backpressure is provided by having each router inform the controller when it transitions in to or out of the state "all-egress-buffers-not-full". The controller will only schedule a broadcast when all egress buffers in all routers are not full. Broadcast backpressure is a configurable option. If it is not activated, the routers do not send status messages and the controller schedules broadcasts on demand. Using this method there is no guarantee that the packet will be forwarded on all ports.

The switching matrix is shown in schematic form in Figure 9. It comprises a high-speed, edge-clocked, synchronous, 16 port dual plane serial cross-point switch SCN for use in the system. It has been optimised to provide a scaleable, high bandwidth, low latency data movement capability. It operates under the control of the controller SM, which sends configuration information to the matrix over the controller interface SMI to create connections for the transmission of data between routers. The buffer and decode logic BDL receives this information and uses it to control the interconnections within the matrix. Data is applied in serial

field to connect ingress and egress ports. For 4 and 8 port configurations, the number of bits of the control port field required is 2 and 3 respectively.

In operation, the switching matrix receives configuration information from the controller SM via the controller interface SMI. This information is loaded into, and stored in, configuration registers. Routing information is passed in the form of a number of encoded fields determining which input port is to be connected to each output port via the switching matrix. In a 16 x 16 matrix, there are 16 output ports. For each output port there is a four bit source address which is encoded to define which input port is to be connected to an output port. There is also an enable signal for each field to signal that the field is valid and a configure signal that indicates that the whole interface is valid. If a field is signalled as not valid, the output port for that field is not connected. If the configure signal is not asserted, the matrix does not change its current configuration. The configuration information on the controller/matrix interface is loaded into the device when the configure signal is asserted. A 16-stage programmable pipeline is used to delay the configuration information until it is required for switching the matrix. If there is a parity error on a port then that ports enable signal will be set to zero and a null tensor will be transmitted to the output of that port. The register that holds the parity error may only be loaded when the configure signal is high and is cleared when read by the diagnostic unit. A parity check is also carried out on the configure signal. If a parity error occurs here then a parity fail condition is asserted, all port enable signals are set to zero and all the output ports on the device will transmit null tensors. The connection between the routers and the matrix is via a set of serial data streams, each running at one Gbaud. Once a connection across the matrix has been set up, tensors are transmitted between ingress and egress routers. The whole process exhibits low latency due to a very small insertion delay. Multiple switching

Alternatively, the matrix is configured according to a probabilistic, work-conserving algorithm located in the priority selector unit PSU.

The router interface unit SIU is provided for every router in the system. Each instance provides the functionality described below. The controller SM monitors the number of tensors in each of the ingress router queues (each router has separate queues for each system destination port, together with a multicast queue, at each of four priority levels). The monitoring is done using a pair of tightly coupled state machines, one in the router and the other in the controller. For small numbers of vectors in a queue, the controller keeps an exact count of the number of vectors. The router notifies the controller when new vectors are added to a queue and the controller decrements the queue size when it schedules one of the vectors in the queue. When there are a larger number of vectors in a queue, the controller keeps only an approximate (fuzzy) count of the queue size and is informed by the router when the queue size crosses predefined boundaries. This minimises the amount of state information that needs to be stored and processed in the controller.

The central management unit CMU is common to all devices. Its functions are to provide the Fabric Management Interface FMI between each device and an external controller, control error management within the device and provide a reset interface RS and reference clocking CK in to each device.

In operation, the controller SM receives requests for connections from the routers over the controller/router interface SRI. As the connection requests arrive, they are queued at the router interface SIU. Since several routers can be requesting connections simultaneously, the controller provides scheduling and arbitration logic to maximise connection efficiency and to ensure that all ports receive a fair

There is also another mechanism in the controller/router interface referred to as 'core level backpressure', which prevents the controller from scheduling any traffic to a particular egress router. A router uses core level backpressure when all its egress buffers are full.

The controllerer is capable of establishing both unicast and broadcast connections in the switching matrix. It is also capable of dealing with system configurations that contain a mixture of 'full' and 'half speed' ports, for example a mixture of 10Gbit/sec and 5Gbit/sec routers.

Figure 12 shows a router device. This is a system port interface control device. Its main function is to support user applications' data movement requirements by providing access into and out of the system. There are two instances of the ingress interface unit IIU, one for each of the traffic managers that can be connected to a system port. The IIU is responsible for transferring data from a traffic manager into an internal FIFO queue on the router and informing the ICU that it has tensors ready to transmit into the system. The external interface to the traffic manager utilises common system interface CSIX. This defines an n x 8-bit data bus; the ingress interface units IIU operate in a 32-bit mode. The FIFO is four tensors deep to allow one tensor to be transferred to the ICU while subsequent ones are being received.

To generate the tensors, the ingress interface unit appends a three byte system core header to the CSIX frame prior to passing it, indicating the tensors availability to the ICU. The IIU examines the CSIX header to determine whether the frame is of type unicast, multicast or broadcast and indicates the type to the ICU. If the frame is unicast, the IIU sets a single bit in byte 1 indicating the destination TM, this is derived from the destination address in the CSIX header. If

statically allocated. There are 512 unicast, 64 multicast and one broadcast queue. The unicast and multicast queues are located in external SRAM. The queue organisation allows flow control down to OC-12 granularity. Within the unicast address field of the CSIX header, 3 bits are allocated for the number of traffic managers a router can support. Since the router supports two traffic managers, the spare bit field is used for a function known as Service Channel. Service Channels provide the means of fully exploiting the routers implicit OC-12 granularity features.

When the ingress control unit ICU receives a connection grant signal from the Controller interface unit SMIU (which specifies egress port and priority), the ICU must choose one of up to 8 qualifying unicast queues or the multicast queue from which to forward a tensor. This is achieved using a weighted round-robin mechanism, that takes into account several parameters. One is the ingress queue length, which allows for the favouring of longer queues over shorter ones and another is aggregate queue tensor urgency, which allows a traffic manager to temporarily increase the weighting of a queue via the urgency field in the CSIX header. One further parameter taken into account is queue bandwidth allocation, whereby an external system controller or system operator can configure the system to provide bandwidth allocation to individual flows via the FMI. The final parameter considered is that of target egress queue backpressure. This requires that the effective performance of the multicast scheme requires that the probability of egress queues being full be minimised. The sensitivity of the weighting function to the input variables is controlled by four sets of global sensitivity variables (one per priority). These settings are configured at system initialisation.

To provide an ingress flow control mechanism, the ingress control unit ICU implements three watermark levels to indicate the state of the queues (fairly empty,

To provide ingress flow control the EIU accepts ingress buffer multistate backpressure information from the ICU and sends it immediately to the traffic mamanger.

The egress control unit ECU is responsible for accepting tensors from the serial transceivers, when informed by the controller interface untit SMIU of their imminent arrival, and forwarding them to the relevant EIU. The ECU examines the traffic manager mask byte of the system core header to determine the correct destination EIU. In the case of multicast (or broadcast) tensors, multiple bits are set in the mask and the tensors are simultaneously transferred to all the EIUs for which a corresponding bit is set. This feature provides wire speed multicasting at the egress router. The ECU is responsible for checking the tensor error check bytes of the system core header. If the system core error checking has been enabled (i.e. the appropriate bit in a status register is set) and the ECU detects an error, then it is logged and the corresponding tensor discarded. To provide an egress flow control mechanism the ECU implements three watermark levels to indicate the state of the egress buffers (fairly empty, filling up, fairly full or very full). When an egress buffer moves from one state to another the ECU signals the change to the ICU. The level of the watermarks is configurable via the FMI. In addition to this multistate backpressure mechanism it is also possible to invoke a second mode of backpressure signalling that involves only start/stop signalling. The type of backpressure mechanism is selected via the FMI by setting the watermark levels appropriately.

The controller interface unit SMIU is responsible for controlling the interface to the controller. Since the controller operates at the system port rather than the traffic manager port level of granularity, the SMIU also operates at this level. The SMIU maintains a count of the number of tensors in the ingress queues

priority in the switch. In addition to the unicast queue there is a multicast queue per port per priority and a single broadcast queue. The unicast and multicast queues are statically allocated in external SRAM. The purpose of this level of buffering is to allow the controller to allocate connections efficiently by giving it a view of the ingress datastreams and to provide rate matching between the router external interfaces and the router/matrix interface.

When connections are granted, the controller creates a connection across the switching matrix to the requested egress router at a given priority. The ingress control unit ICU must now choose one of the qualifying unicast or multicast queues from which to forward a tensor to the transceiver for serialisation. This level of router scheduling is done on a weighted-round-robin-basis. Each unicast and multicast queue has weighting associated with it, which is determined by the backpressure from the egress buffers, the queue length, the queue urgency and the static bandwidth allocation. On the egress side the controllerer informs the router of a tensor's imminent arrival. The egress control unit ECU receives this tensor and examines the core header to see which traffic manager to send the tensor to. Tensors are then assembled back into datastreams and forwarded via CSIX to the appropriate traffic manager.

Multicasting in the system is achieved by the optimal replication of tensors at the ingress and egress. On the ingress side a router has one multicast queue per egress router at each priority. Multicast routing information is appended on the ingress side and on arrival at the egress side these masks determine the replication of tensors into the required egress buffers. Broadcast in the system is achieved by having a single on chip broadcast queue at the ingress of each router. When the controller schedules a broadcast connection, the tensor will be routed by the matrix to all egress routers in parallel, thus avoiding any ingress congestion.

Each system device contains a logic block known as the fabric management interface unit (FMIU). The FMIU interfaces to the functional logic, also known as the core, within the device in order to provide run-time (read/write) access to a chosen subset of the registers and RAM locations, a mechanism to report run-time fail conditions detected by the device, and scan access (read/write) to the total set of registers in the functional logic while the functional logic is not operational.

The external interface to the fabric management interface unit FMIU requires a number of inputs, including a Hard Reset input which sets the system device into a known state. In particular, it sets the device into a state where the FMIU is fully functional and the serial interface can be used. Hard Reset is expected to be applied when power is first applied to the device, and may also be applied at other times. The external interface also has a serial inpt and serial output lines and a device locator address field used to identify a particular instance of a device. The device locator field is generated by tie-offs that are determined by the devices physical position in the system.

The main functions of the central management unit (CMU) shown in Figure 12 include error detection and logging logic. This is responsible for detecting error conditions and states within the chip or on its interfaces. As such, its functionality is spread throughout the design and is not concentrated within a specific block. Errors are reported and stored in the Error and Status registers and logs, which are accessible across the FMI. The CMU also has reset and clock generation logic responsible for the generation and distribution of clocks and reset signals within the device. In addition, the CMU contains test control logic which controls the mechanisms built in for chip test. The target fault coverage is 99.9%. This logic is not used under normal operating conditions. The final function of the CMU is to provide fabric management logic common to all of the system devices

CLAIMS

1. A method of handling packets of information through a data switch comprising input traffic controllers, ingress routers, a memoryless cyclic switch fabric, egress routers and output traffic controllers all under the control of a switch controller and interconnected such that each input line connected to the data switch is terminated on a traffic controller arranged to convert the input line protocol information packets into fixed length cells having a header defining the data switch destination router and output traffic controller together with message priority information arranged such that each ingress router serves a group of traffic controllers characterised in that the ingress router includes a set of input buffers one for each input line and a set of virtual output queue buffers, one for each output traffic controller from the data switch, and in which the method comprises on the arrival of a cell from a traffic controller the ingress router examines the cell header and places it in the appropriate virtual output queue and generates a request for transfer message consisting of the destination traffic controller address and a message priority code which is passed to the data switch controller, the switch controller schedules the passage of the cells across the switch fabric by interconnecting a specific ingress router to a specific egress router for each switch fabric cycle in accordance with a first arbitration process the ingress router selecting from the appropriate virtual output queue the cell at the head of the queue for passage across the data switch to the appropriate output traffic controller in accordance with a second arbitration process.

2. A method of handling packets of information through a data switch as claimed in claim 1 characterised in that the ingress buffering is organised into separate queues, one for each destination traffic controller and each priority level.

router to a specific egress router for each switch fabric cycle in accordance with a first arbitration process and the ingress router selects from the appropriate virtual output queue the cell at the head of the queue for passage across the data switch to the appropriate output traffic controller in accordance with a second arbitration process.

6. A data switch for handling packets of information as claimed in claim 5 characterised in that the virtual output queues are arranged as separate queues one for each destination traffic controller and each priority level.

7. A data switch for handling packets of information as claimed in claim 5 or 6 characterised in that the ingress router uses a weighted round-robin mechanism to select the next queue buffer based on ingress queue length, aggregate queue packet urgency and target traffic controller egress queue backpressure.

8. A data switch for handling packets of information as claimed in claim 5, 6 or 7 characterised in that the switch controller performs a first arbitration process which involves determining the set of requests to be accepted for each switch fabric cycle by attempting to deliver a packet of information to each output switch fabric port in every arbitration cycle.

9. A method of handling packets of information through a switch as described and shown in the accompanying drawings.

10. A data switch for handling packets of information as described and shown in the accompanying drawings.

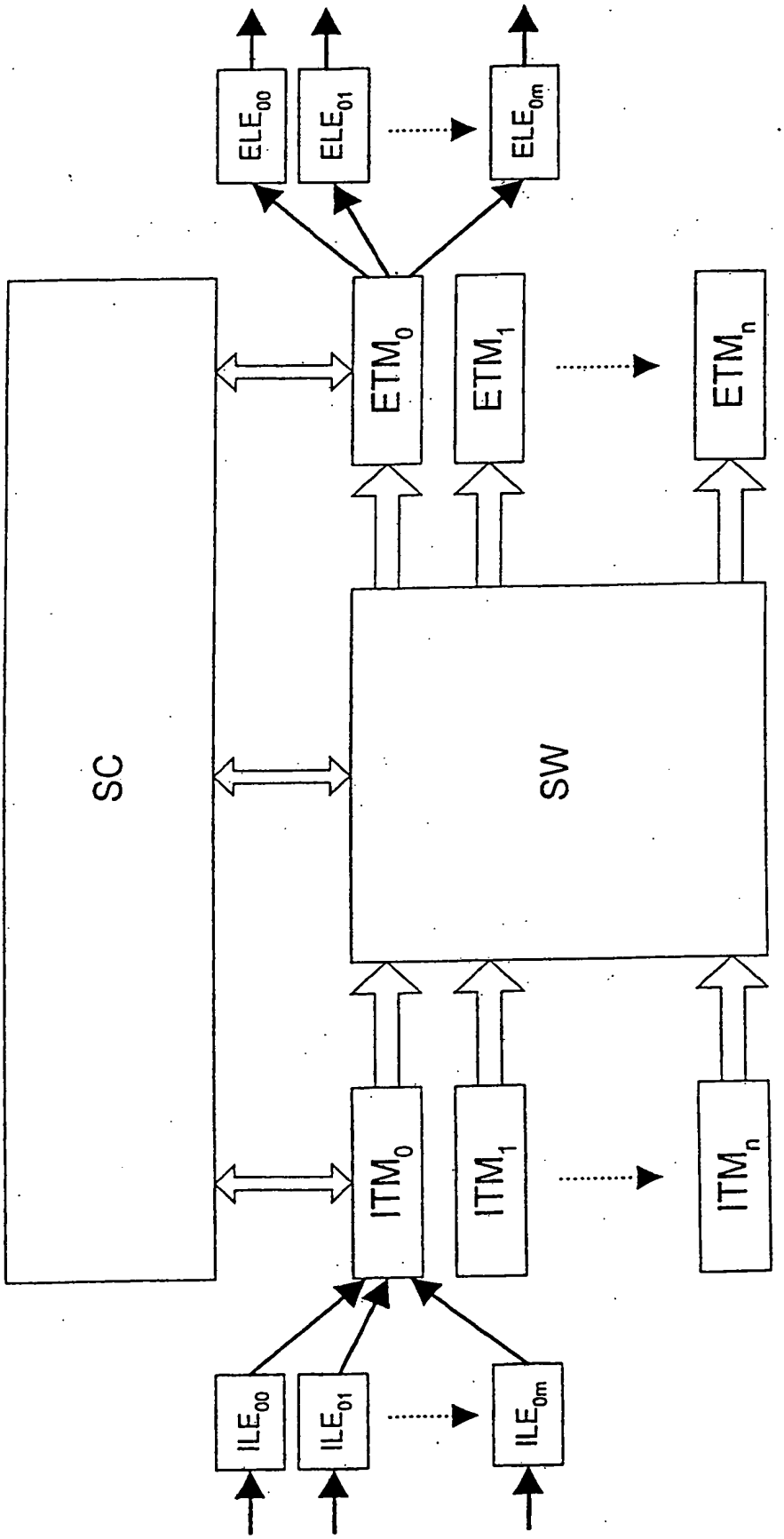

ELE00
ELE01
ELE0m
ETM0
ETM1
ETMn
SC
SW
ITM0
ITM1
ITMn
ILE00
ILE01
ILE0m

Fig 2.

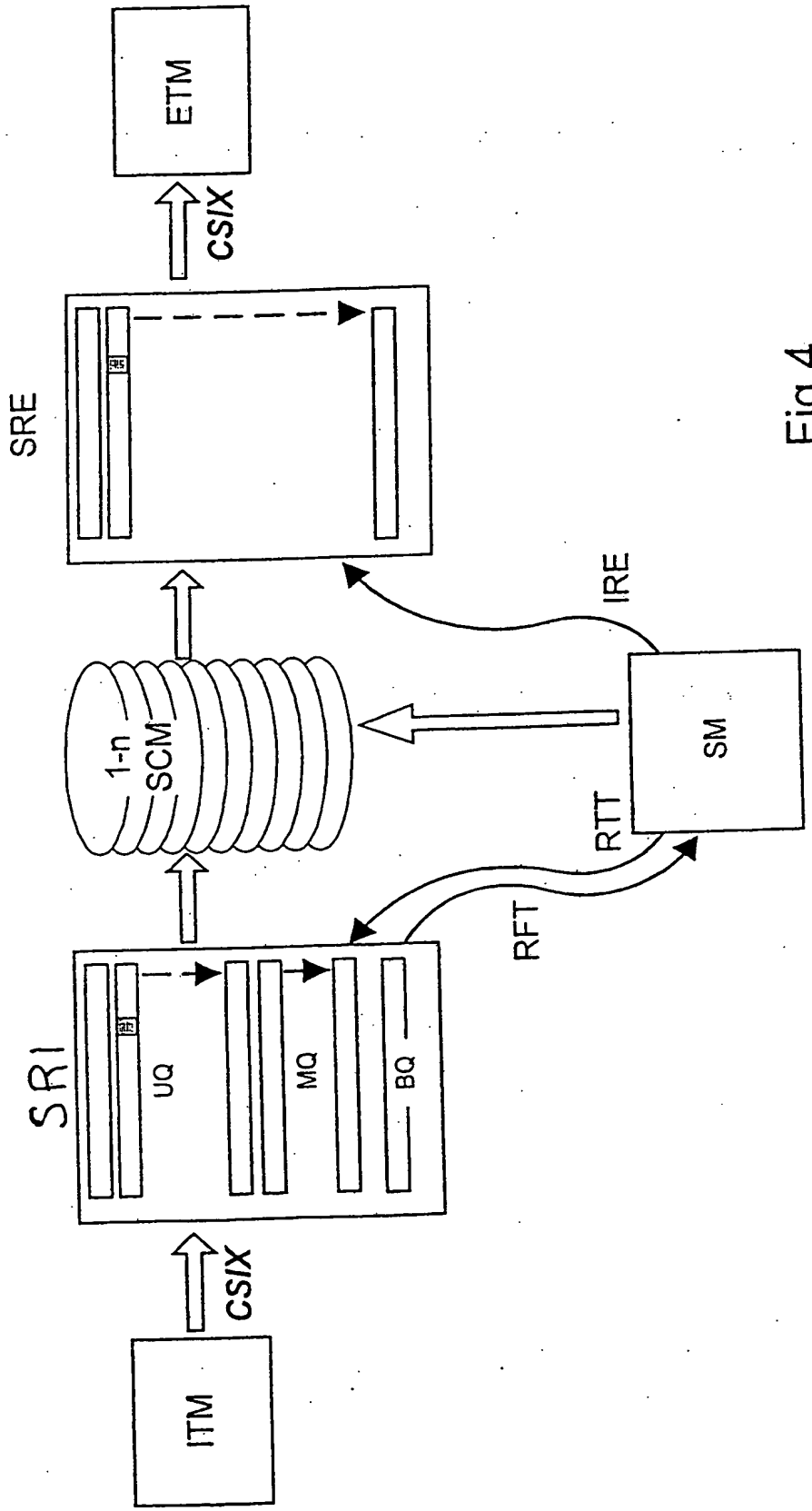


Fig 4.

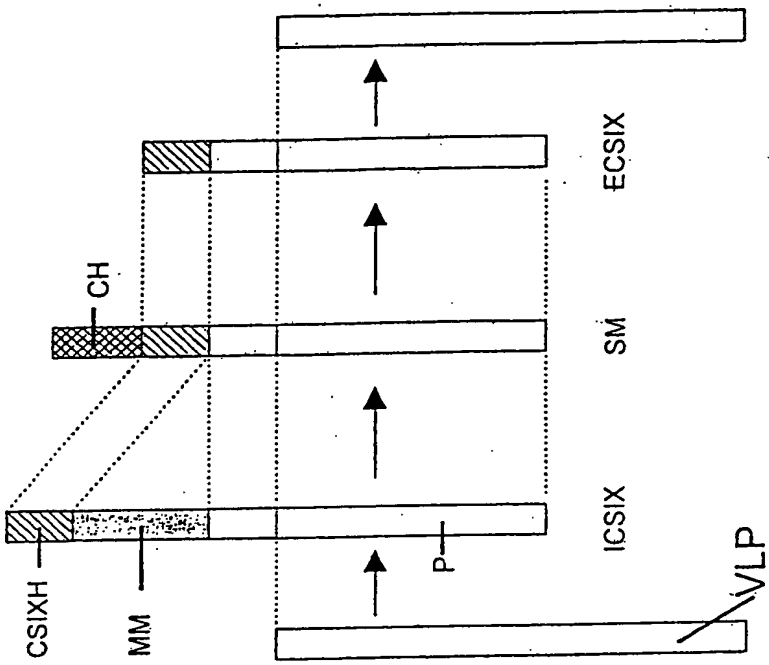
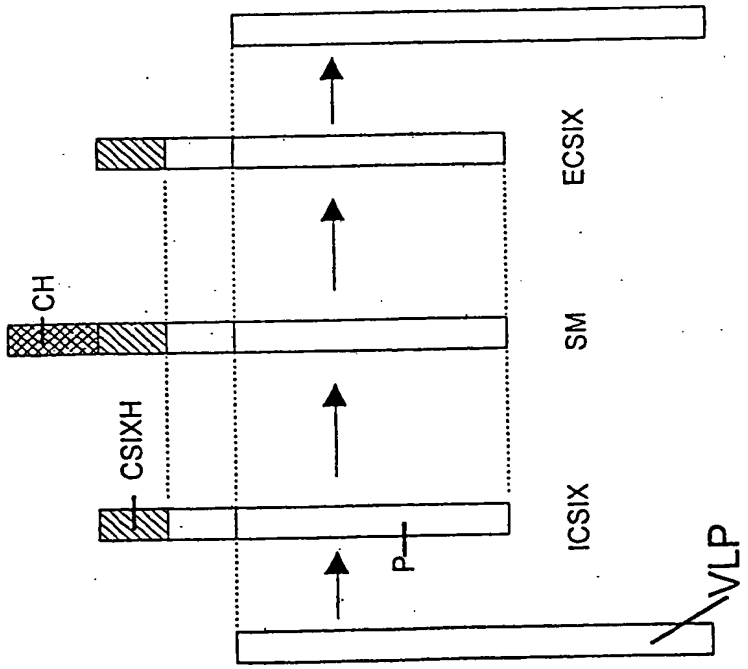


Fig 6.



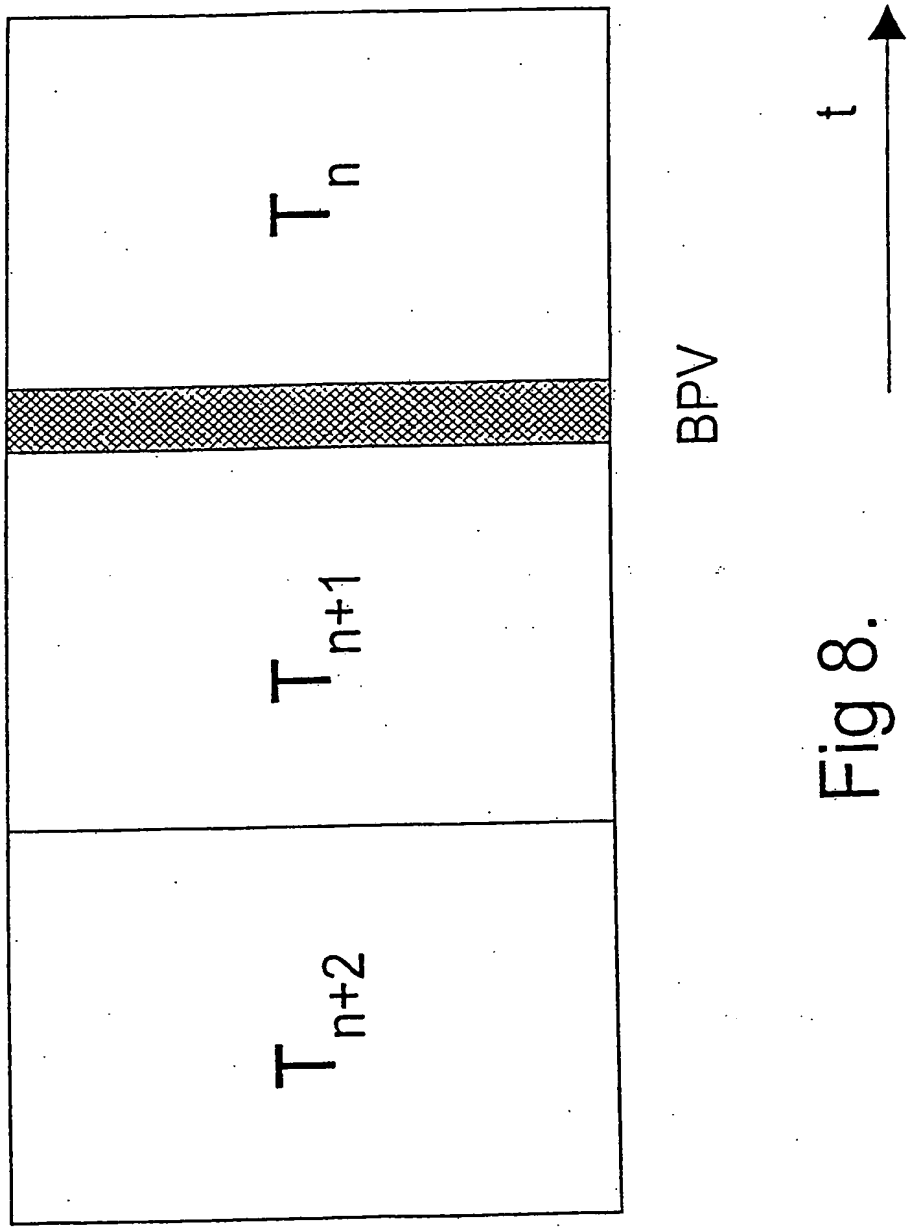


Fig 8.

Fig 10.

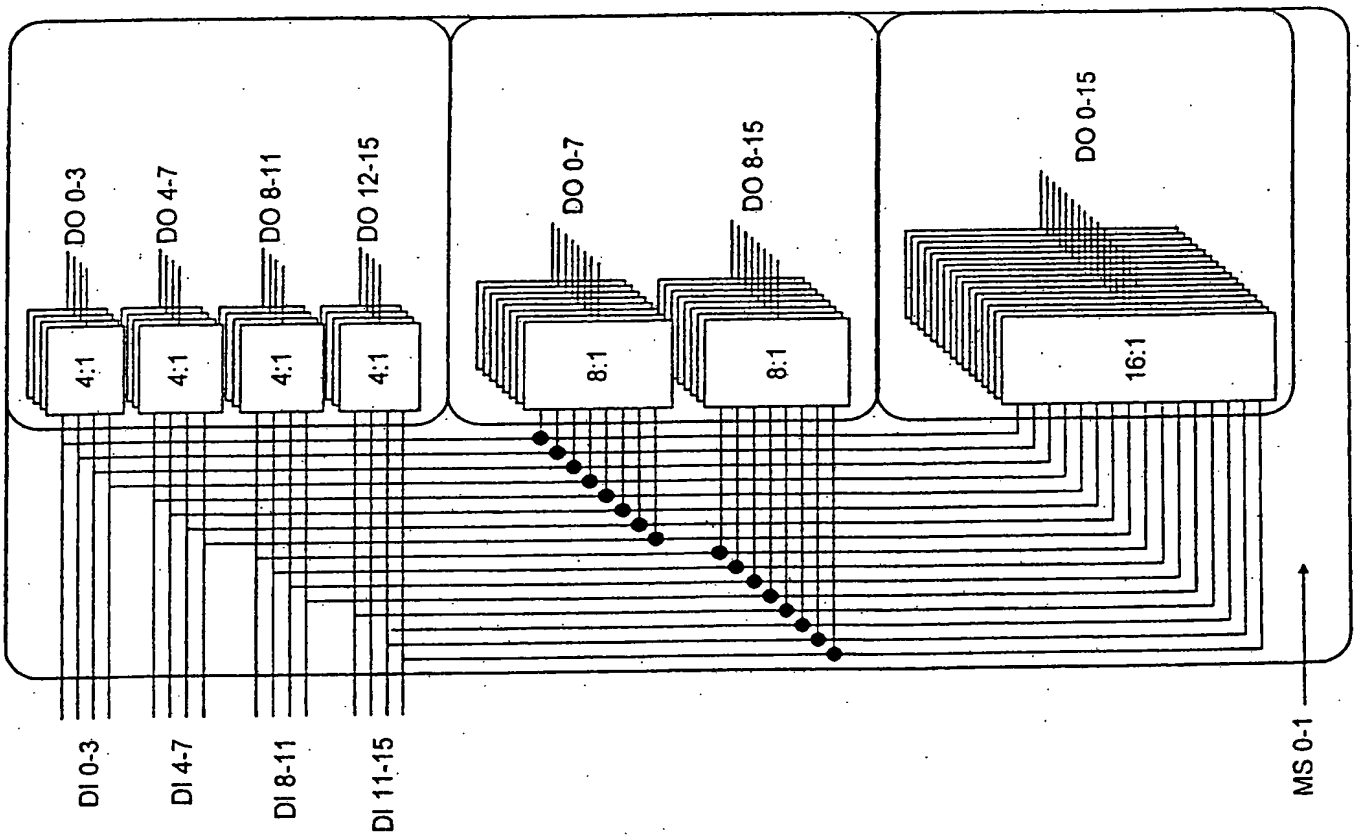
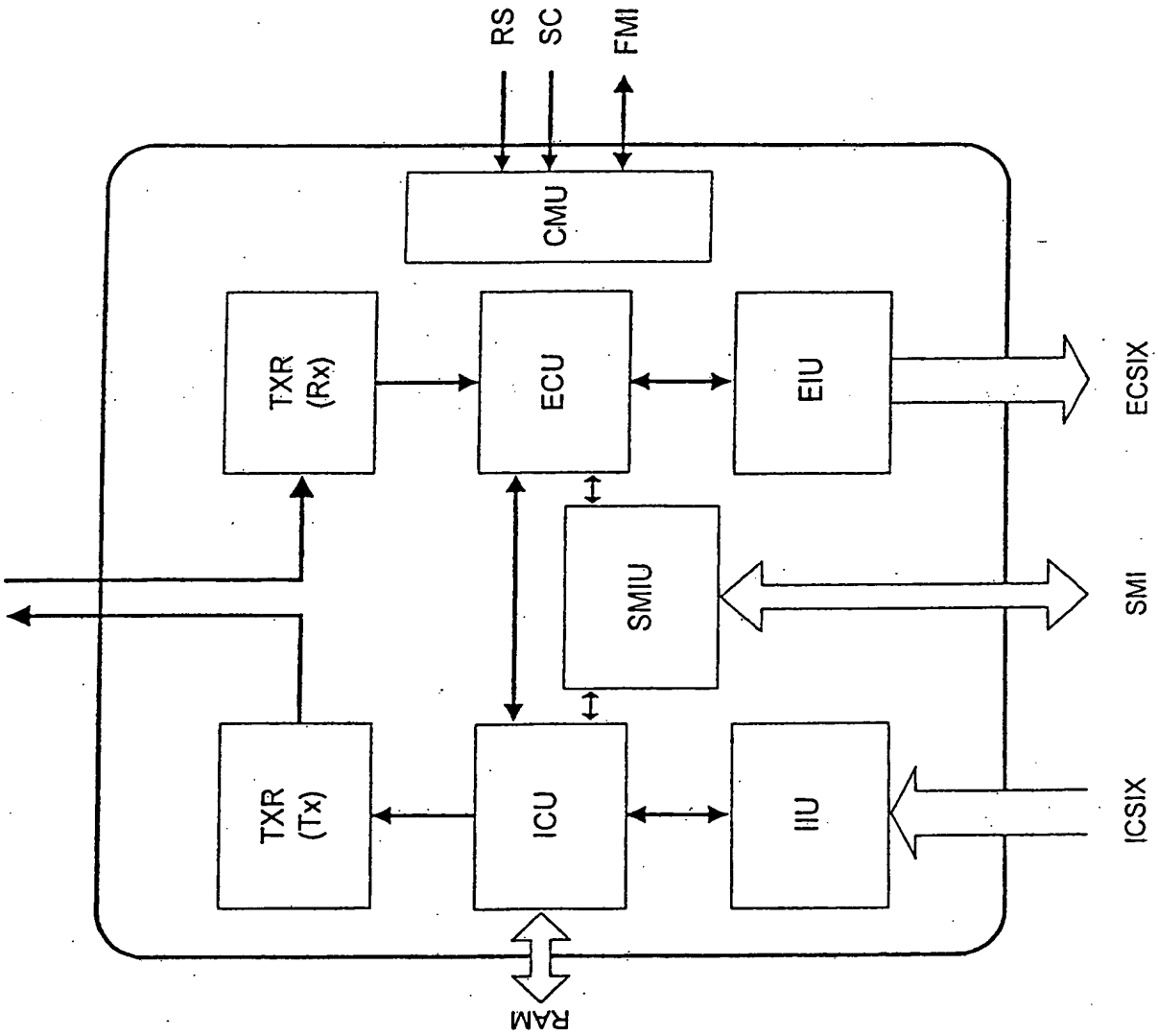

DO 0-3
DO 4-7
DO 8-11
DO 12-15
DO 0-7
DO 8-15
DO 0-15
4:1
4:1
4:1
4:1
8:1
8:1
16:1
DI 0-3
DI 4-7
DI 8-11
DI 11-15
MS 0-1

Fig 12.


RS
SC
FMI
CMU
TXR (Rx)
ECU
EIU
ECSIX
SMIU
SMI
TXR (Tx)
ICU
IIU
ICSIX
RAM

INTERNATIONAL SEARCH REPORT

In ational Application No
PCT/GB 99/03748

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 H04L12/56 H04L12/64 H04Q11/04

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 H04L H04Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 0 849 916 A (IBM) 24 June 1998 (1998-06-24) abstract ---	1,5
A	MCKEOWN N ET AL: "TINY TERA: A PACKET SWITCH CORE" IEEE MICRO,US,IEEE INC. NEW YORK, vol. 17, no. 1, 1 January 1997 (1997-01-01), pages 26-33, XP000642693 ISSN: 0272-1732 page 27, left-hand column, line 1 -page 28, right-hand column, line 27 -----	1,5

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 February 2000	25/02/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Dhondt, E

Form PCT/ISA/210 (second sheet) (July 1992)

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☑ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.